PLoS one

# Phylogenomics: Gene Duplication, Unrecognized Paralogy and Outgroup Choice

**Scott William Roy***

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

## Abstract

Comparative genomics has revealed the ubiquity of gene and genome duplication and subsequent gene loss. In the case of gene duplication and subsequent loss, gene trees can differ from species trees, thus frequent gene duplication poses a challenge for reconstruction of species relationships. Here I address the case of multi-gene sets of putative orthologs that include some unrecognized paralogs due to ancestral gene duplication, and ask how outgroups should best be chosen to reduce the degree of non-species tree (NST) signal. Consideration of expected internal branch lengths supports several conclusions: (i) when a single outgroup is used, the degree of NST signal arising from gene duplication is either independent of outgroup choice, or is minimized by use of a maximally closely related post-duplication (MCRPD) outgroup; (ii) when two outgroups are used, NST signal is minimized by using one MCRPD outgroup, while the position of the second outgroup is of lesser importance; and (iii) when two outgroups are used, the ability to detect gene trees that are inconsistent with known aspects of the species tree is maximized by use of one MCRPD, and is either independent of the position of the second outgroup, or is maximized for a more distantly related second outgroup. Overall, these results generalize the utility of closely-related outgroups for phylogenetic analysis.

**Citation:** Roy SW (2009) Phylogenomics: Gene Duplication, Unrecognized Paralogy and Outgroup Choice. PLoS ONE 4(2): e4568. doi:10.1371/journal.pone.0004568

**Editor:** Niyaz Ahmed, University of Hyderabad, India

**Received** September 18, 2008; **Accepted** November 27, 2008; **Published** February 23, 2009

**Funding:** National Library of Medicine Internal Research Training Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The author has declared that no competing interests exist.

* E-mail: royscott@ncbi.nlm.nih.gov

## Introduction

Accurate phylogenetic inference is thwarted by the presence of conflicting signals in the data (e.g., [1–3]), a problem that has received a large amount of theoretical and experimental attention over the past few years (e.g., [4–7]). One important general source of problems in phylogenetic analysis is gene duplication (e.g., [8–10]).

In particular, loss of different members of an ancestral duplicate pair in different species can lead to a gene tree that does not reflect the species tree [11–13]. Recent studies in yeast and teleost fish [14–16] suggest that such reciprocal loss of gene duplicates following genome duplication may be a common phenomenon, raising the specter of significant conflicts between gene trees and species trees in such lineages. Such cases are particularly troublesome since they are expected to pass the most common bioinformatics test for one-to-one orthologs: genes from different species that are each other's best reciprocal BLAST hit. While under such circumstances every attempt to identify and eliminate such paralogs from analyses should be made, it remains likely that some paralogs will persist in many-locus sets of putative orthologs. Given the importance of gene duplication in the evolution of eukaryotic genomes in general and of several lineages of great interest in particular [17–19], understanding such challenges is useful for correct reconstruction of evolutionary history.

Recently, Rogozin et al. [20] suggested that in the case of whole-genome duplication, use of a closely related outgroup – specifically one that diverged from the species of introns following the duplication – might lead to incorrect species tree reconstruction. Thus, in stark contrast to the general case in the absence of gene duplication, use of a more distant outgroup might be more likely to yield an accurate tree. A great deal of theoretical and empirical work has vindicated taxonomic sampling to reduce branch lengths, including use of closely related outgroups [21–23]; however the case of multi-locus analyses in the presence of differential loss of ancestral gene duplicates has not been addressed to my knowledge. Here I address the issue of outgroup choice in the presence of NST signal arising from subsets of unrecognized paralogs genes within many-gene sets of putative orthologs.

## Analysis and Discussion

### Phylogenetic reconstruction in the presence of unrecognized paralogs

I consider the problem of resolving the relationship between three species/groups by studying many-gene sets of putative one-to-one orthologs, in which some genes have been duplicated in the ancestor of the three species/groups of interest. Following such an ancestral duplication, there are three possible outcomes. Some studied species might retain both duplicates, in which case the gene will not be a one-to-one ortholog, and thus should be recognized and discarded. If not, the two most closely related species/groups might retain the same duplicate copy (either by independent loss of the other duplicate, or loss of one duplicate in an ancestor of the two species/groups), in which case the topologies of the gene and species trees will be identical, and thus the gene duplication is not expected to pose problems to reconstruction of the species tree. Finally, all three species/groups might lose one duplicate copy, but the two most closely related groups lose different (reciprocal) copies, in which case the species and gene trees will have different topologies. It is that third case that I address here.

The problem is treated under the following conditions and assumptions. First, I consider phylogenetic analysis of concatenated alignments across putative orthologs. Notably, the present arguments do not apply for methods such as matrix representation parsimony, in which trees are reconstructed from individual gene alignments. In such cases the length of the branch supporting the incorrect (non-species) grouping is not of consequence, but only the topology itself. Second, I assume that in general shorter NST internal branches are preferable to long ones, since such branches are expected to generally experience fewer changes and thus contribute less NST signal.

## Gene and species trees under gene duplication and loss

Gene duplication and differential subsequent gene loss can lead to gene trees that do not reflect the species tree. The greater the strength of this alternative (NST) signal, the greater the chance of recovering the wrong tree in multi-locus analyses. The general case is illustrated in Figure 1a. Species C represents an outgroup to species A and B. A gene duplication (grey diamond) occurred at a time *d* before the C/AB split. Both duplicates were retained all the way to the A/B split. The three species then each lost one of the two duplicates (dotted grey lines), returning to a single-copy state (solid black lines). Although species C represents an outgroup to species A and B, retention of the same duplicate copy in species A and C (left side) but the other copy in species B (right side) leads to a case in which the remaining copies in A and C are most closely related. (Note that, throughout, all figures consider the case in which C and A retain the same duplicate. For each case there is an equivalent case in which C and B (but not A) retain the same duplicate).

## One outgroup

I next consider gene trees with a single outgroup. Such an outgroup may diverge from the ABC ancestor before the duplication (Figure 1b) or may diverge after the duplication and subsequently lose one duplicate copy (Figure 1c–e; if the outgroup retains both duplicates, the gene will be recognized as not a one-to-one ortholog, and should be discarded). Figure 1c gives the general scenario, and Figure 1d and e illustrate the cases in which either duplicate is lost in species X, with both rooted and simplified unrooted trees shown.

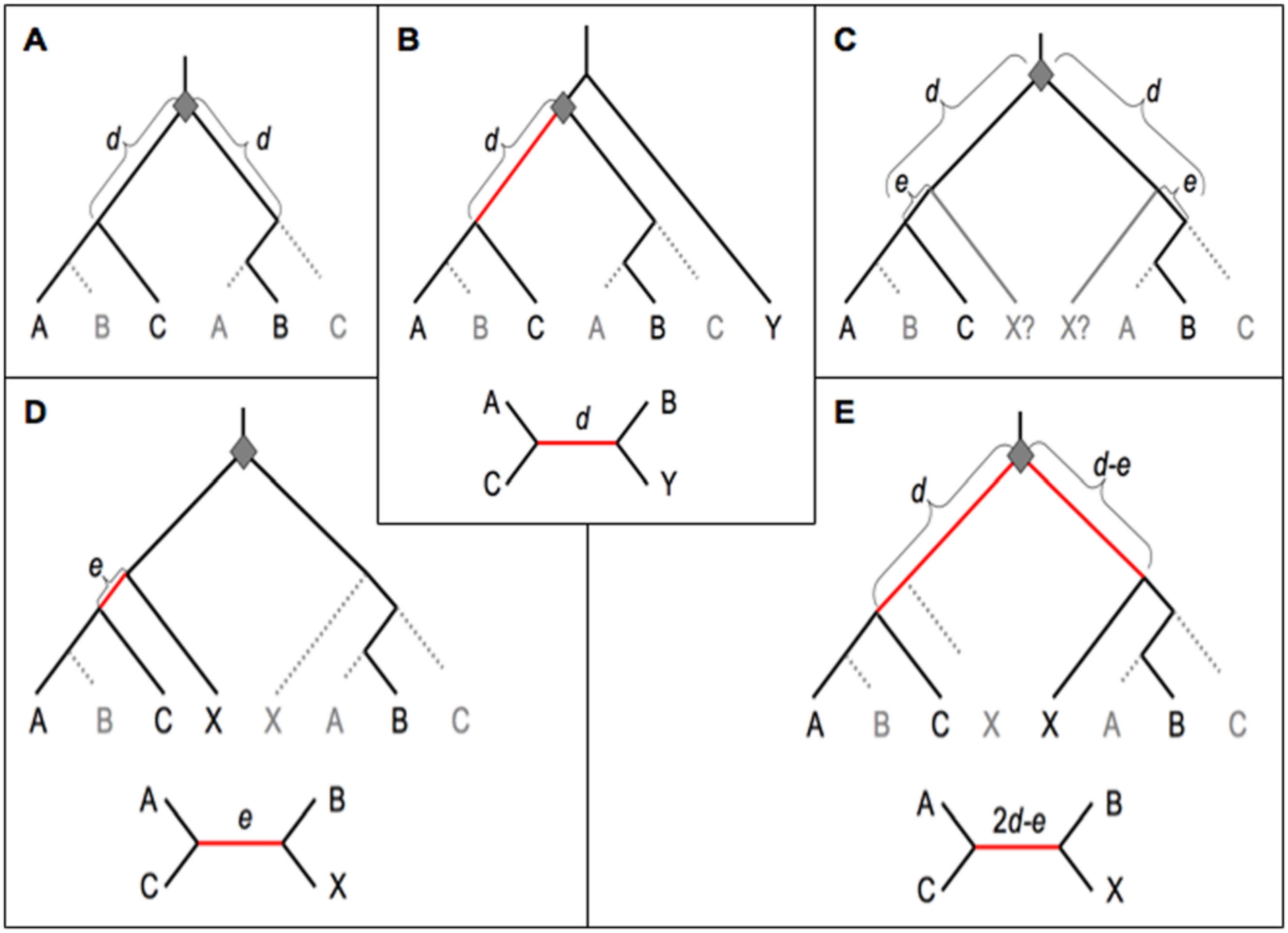

**Figure 1. Differential loss of ancestral duplicates and gene trees with single outgroups, for different phylogenetic positions of the outgroup.** a) The general case in which duplication (grey diamond) followed by differential gene loss (dotted grey lines) leads to a closer relationship between gene copies from species A and C, in contrast to the species relationship. b) The case for a single pre-duplication outgroup Y. c) The general case for a post-duplication outgroup X diverged time *e* before the C/AB split. d) X and C retain the same duplicate copy. e) X and B retain the same duplicate copy. Both rooted and unrooted trees are shown. Red branches indicate the NST internal branch, supporting an A+C group. doi:10.1371/journal.pone.0004568.g001

As Figure 1b,d,e shows, the four-taxa gene tree consistently contradicts the species tree regardless of the position of the outgroup relative to the duplication event (pre- or post-) and regardless of which duplicate is lost in a post-duplication outgroup. That is, in each case the gene tree contains an A+C group (red branches), in contrast to the true species grouping A+B. How does outgroup position affect the amount of NST signal? NST signal is due to changes occurring along the aberrant internal branch (red in Figures 1, 2, 3). The shorter this internal branch, the fewer expected genetic changes and the less NST signal, so in general outgroups that minimize NST internal branch length are preferable.

Using a pre-duplication outgroup (Figure 1b), the length of the NST internal branch is *d*. Using a post-duplication outgroup the

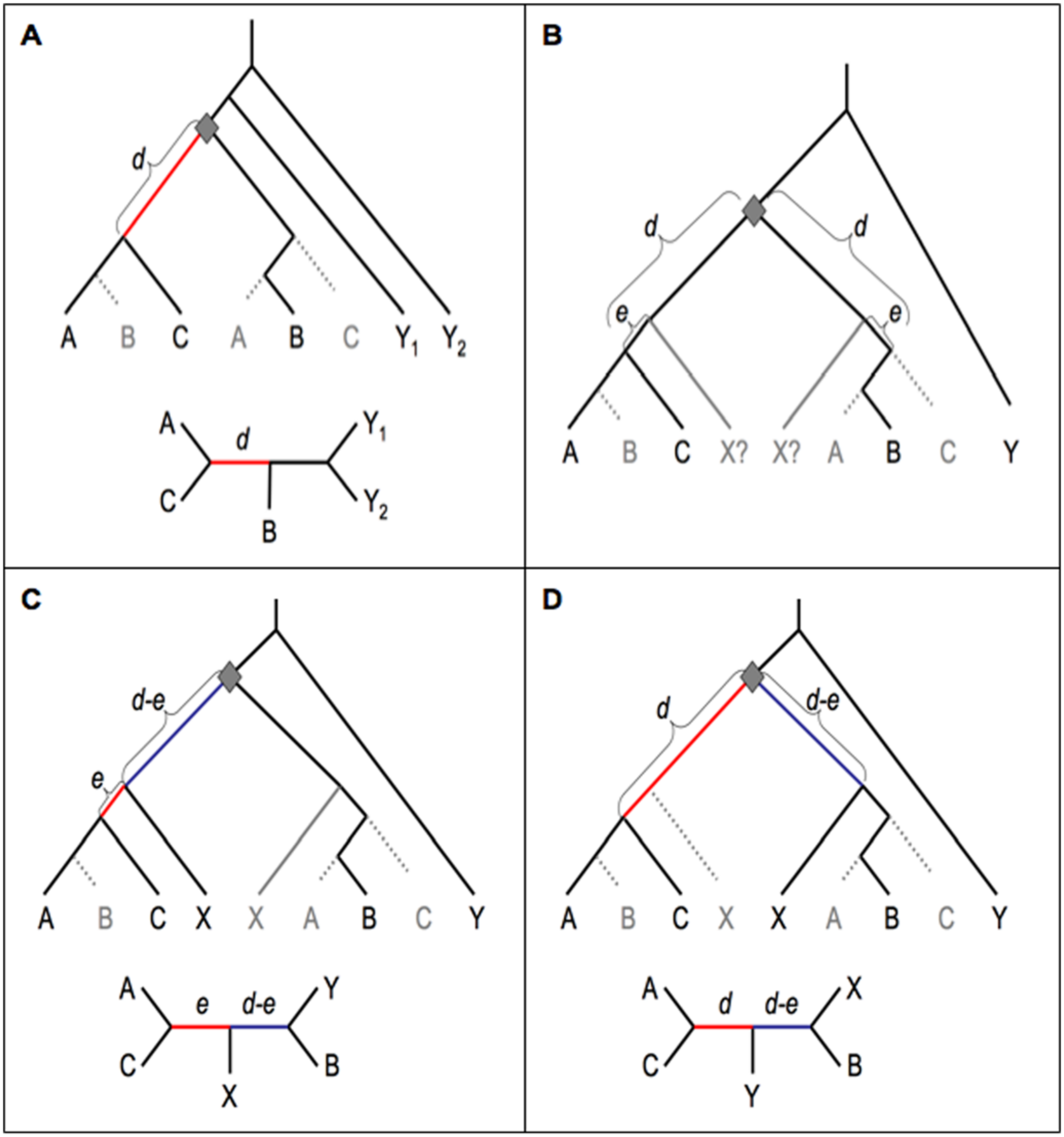

**Figure 2. Gene trees with two outgroups, including at least one pre-duplication outgroup.** a) Two pre-duplication outgroups $Y_1$ and $Y_2$. b) The general case for one pre-duplication outgroup Y and one post-duplication outgroup X. c) X and C retain the same duplicate copy. d) X and B retain the same duplicate copy. Red branches indicate the NST internal branch supporting an A+C clade. Blue branches indicate the known non-species tree (KNST) branch, contradicting the known A+B+C group.
doi:10.1371/journal.pone.0004568.g002

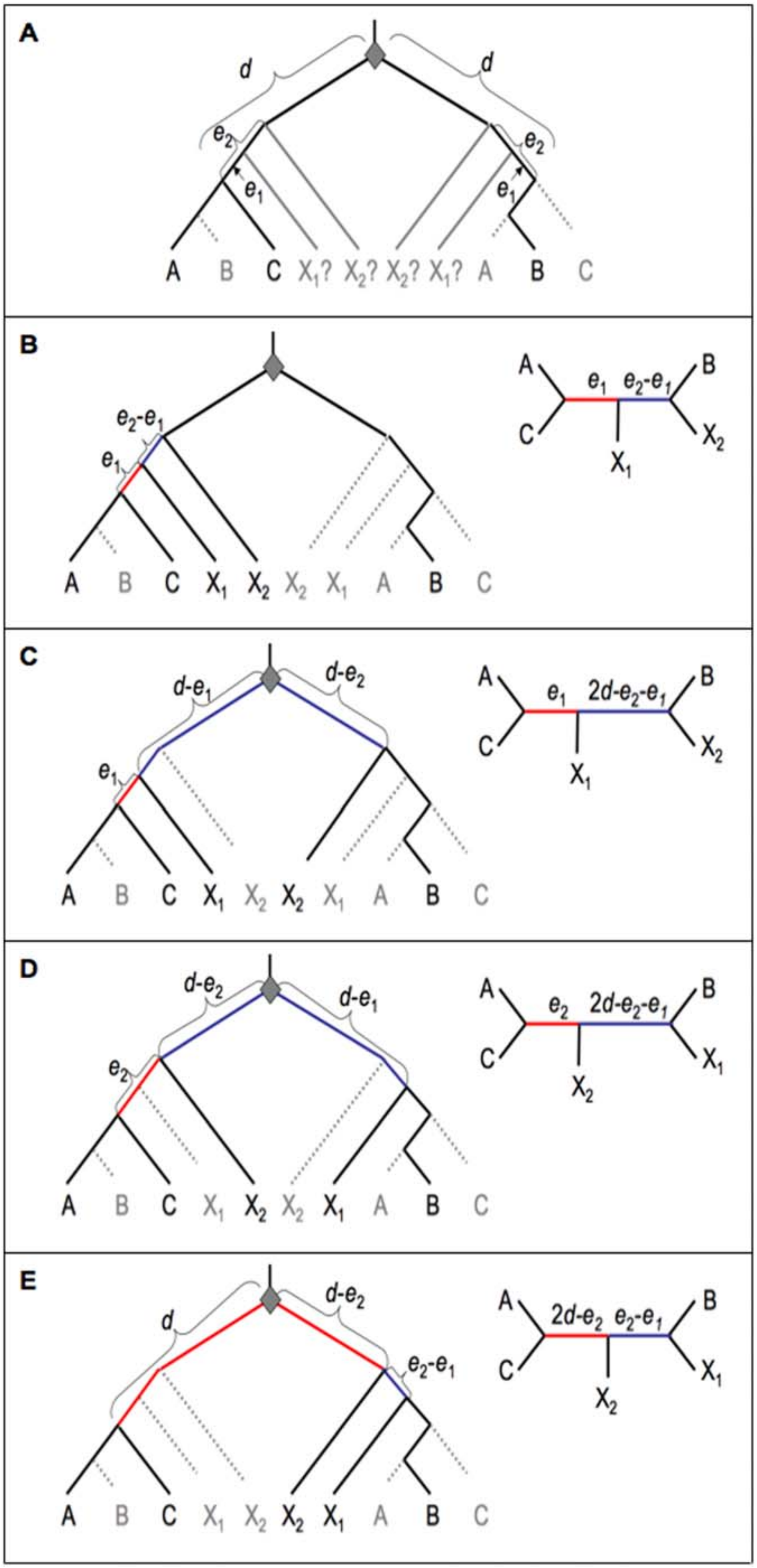
A
d
d
$e_2$
$e_2$
$e_1$
$e_1$
A B C $X_1$? $X_2$? $X_2$? $X_1$? A B C
B
$e_2$-$e_1$
$e_1$
A B C $X_1$ $X_2$ $X_2$ $X_1$ A B C
A C $e_1$ $e_2$-$e_1$ B $X_2$ $X_1$
C
$d$-$e_1$
$d$-$e_2$
$e_1$
A B C $X_1$ $X_2$ $X_2$ $X_1$ A B C
A C $e_1$ $2d$-$e_2$-$e_1$ B $X_2$ $X_1$
D
$d$-$e_2$
$d$-$e_1$
$e_2$
A B C $X_1$ $X_2$ $X_2$ $X_1$ A B C
A C $e_2$ $2d$-$e_2$-$e_1$ B $X_1$ $X_2$
E
d
$d$-$e_2$
$e_2$-$e_1$
A B C $X_1$ $X_2$ $X_2$ $X_1$ A B C
A C $2d$-$e_2$ $e_2$-$e_1$ B $X_1$ $X_2$

**Figure 3. Gene trees with two post-duplication outgroups.** a) The general case for two post-duplication outgroups $X_1$ and $X_2$, diverged $e_1$ and $e_2$ before that C/AB split, respectively. b) $X_1$, $X_2$ and C retain the same duplicate copy. c) $X_1$ and C (but not $X_2$) retain the same duplicate copy. d) $X_2$ and C (but not $X_1$) retain the same duplicate copy. e) $X_1$, $X_2$ and B retain the same duplicate copy. Red and blue indicate NST and KNST branches, respectively.
doi:10.1371/journal.pone.0004568.g003

internal branch length is either $e$, the time between the X/ABC and C/AB divergences (if X and C retain the same duplicate copy; Figure 1d) or $2d$-$e$ (if X and C retain different copies; Figure 1e). If these two possibilities are equally likely, the average branch length is $d$, equal to the value for a pre-duplication outgroup.

What if one copy is more likely to be retained? In this case, it is more likely that X and C retain the same duplicate, and the expected NST internal branch length will be less than $d$. In the case where the two duplicates' probabilities of retention differ by a factor $r$, the two duplicates have probabilities $0.5(1\pm r)$ along each branch, and, assuming independent loss along each branch, the probability that X and C retain the same duplicate is $[0.5(1+r)]^2+[0.5(1-r)]^2=0.5(1+r^2)$, giving an average internal branch length of $d\text{-}r^2(d\text{-}e)$, which is less than $d$. Thus in this case: (i) a post-duplication outgroup yields shorter average internal branches, and (ii) later-diverging post-duplication outgroups yield shorter internal branches than do more distant ones (i.e. small values of $e$).

In total then, in the case of a single outgroup, a post-duplication outgroup is a better choice for several reasons: (i) the outgroup may retain both duplicates, allowing detection of the otherwise cryptic ancestral gene duplication; (ii) the well-known general advantages of using closely related outgroups; and (iii) lesser or equal NST signal issuing from cases of differential loss of ancestral duplicates.

## Two outgroups

How do these considerations change when two outgroups are used? In this case, zero, one, or two outgroups may predate the duplication event (Figure 2,3). If both outgroups predate the duplication event, the average NST internal branch length is again $d$ (Figure 2a). With one pre- and one post-duplication outgroup (Figure 2b), the internal branch length is either $e$ (if X and C retain the same copy; Figure 2c) or $d$ (if X and C retain different copies; Figure 2d). Thus in this case the average internal branch length is $e$ X $0.5(1+r^2)$+$d$ X $0.5(1-r^2)=0.5((d+e)-r^2(d-e))$. This is less than or equal to $d$ (since $e \leq d$), thus (i) mixed pre/post duplication outgroups are expected to lead to less NST signal than are two pre-duplication outgroups, and (ii) a more closely related post-duplication outgroup is expected to lead to less NST signal than for a more distant one (i.e., small $e$ value).

The case is somewhat more complicated for two post-duplication outgroups. Figure 3a illustrates the general case, with outgroups $X_1$ and $X_2$ that diverged at times $e_1$ and $e_2$ before the C/AB divergence, respectively (with $e_1<e_2<d$). Now there are four possibilities based on which duplicate copies are retained by $X_1$ and $X_2$ relative to C. If $X_1$ and C retain the same copy (with probability $0.5(1+r^2)$), the internal NST branch is $e_1$ (regardless of which duplicate is retained in $X_2$; Figure 3b,c). If $X_1$ and C retain different copies (with probability $0.5(1-r^2)$), the issue becomes more complicated still. In this case, the NST internal branch is $e_2$ if $X_2$ and C retain the same copy (Figure 3d), or $2d$-$e_2$ if $X_2$ and C retain different copies (Figure 3e).

Which of these two scenarios (i.e. Figure 3d and e) is more likely? The probability of $X_2$ retaining the same copy as C is equal to the probability that $X_2$ and C (but not $X_1$) retain the most likely duplicate, $0.5(1+r)$ X $0.5(1+r)$ X $0.5(1-r)$, plus the probability that $X_2$ and C (but not $X_1$) retain the less likely duplicate, $0.5(1-r)$ X $0.5(1-r)$ X $0.5(1+r)$. This sum is simply $0.25(1-r^2)$. The probability of $X_2$ and $X_1$ (but not C) retaining the same copy is identical, following the same reasoning. Thus the scenarios in Figure 3d and 3e have equally probability. The total expected NST internal branch length is then equal to (i) the probability that $X_1$ and C retain the same duplicate, $0.5(1+r^2)$, times $e_1$, plus (ii) the probability that $X_1$ and C retain different duplicates, $0.5(1-r^2)$, times the average of $e_2$ and $2d$-$e_2$ (which is $d$). This yields a total average expected NST internal branch length of $0.5((d+e_1)-r^2(d-e_1))$. This value is equal to the expectation for one post-duplication outgroup diverging at time $e_1$ and one pre-duplication outgroup (as shown above). Thus, using multiple outgroups, NST signal is minimized by using one closely-related post-duplication outgroup, while the position of the second outgroup (whether pre-duplication, or diverging any time between the duplication and C/AB divergence) does not affect the degree of NST signal in this model.

## Maximizing signal for gene tree branches contradicting known species relationships

In the case of multiple outgroups, a second consideration comes into play. If the additional species (Y, $X_1$, $X_2$) are confidently known to be true outgroups to an A+B+C clade, then the possibility exists for detecting gene trees that contradict known species relationship, allowing for exclusion of such suspect genes from multi-locus analyses. For instance, for the mixed pre/post duplication outgroup case shown in Figure 2c, the gene tree contains a known non-species tree (KNST) branch (shown in blue) dividing A+C+X from B+Y, which contradicts the known A+B+C group. Since this contradictory signal is useful in detecting suspicious gene trees, optimal outgroup combinations will maximize, not minimize, the length of the KNST.

Firstly, and most clearly, the presence/absence of a KNST branch differs between different outgroup combinations. For two pre-duplication outgroups, the gene tree still contains an A+B+C group (Figure 2a), thus there is no KNST branch. By contrast, all cases with at least one post-duplication outgroup (Figure 2c,d and Figure 3b–e, excepting post-duplication pairs for which $e_1=e_2$) contain a KNST branch, thus outgroup combinations that include at least one post-duplication outgroup are preferred.

For mixed pre/post duplication outgroup pairs the length of the KNST branch is $d$-$e_1$ (Figure 2c,d) regardless of which duplicate is retained in species X. This value is maximized for small $e_1$, again supporting usage of a closely related outgroup. For two post-duplication outgroups, the KNST branch length is $e_2$-$e_1$ (if both outgroups retain the same duplicate; Figure 3b,e) or $2d$-$e_2$-$e_1$ (if different duplicates are retained; Figure 3c,d). Thus if both duplicates are equally likely to be retained, the average KNST branch length is $d$-$e_1$, the same as for mixed pre/post duplication outgroups. If instead duplicates are not equally likely to be retained, the expected average length is smaller than for mixed outgroups: $0.5(1+r^2)(e_2\text{-}e_1)+0.5(1-r^2)(2\text{d-}e_2\text{-}e_1)=d\text{-}e_1\ \text{-}\ r^2(d\text{-}e_2)$. In this case, KNST branch length is maximized when one outgroup divergence is maximally close (i.e. coincident with the C/AB divergence: $e_1=0$) and the other outgroup is coincident with the duplication ($d=e_2$). Thus, under unequal probabilities of retention, KNST branch length is maximized for one very close outgroup and one more distant one. As such, depending on the relative probabilities of retention of the two duplicates, the ability to detect aberrant internal branches may provide an exception to the generally better expected performance of maximally closely-related outgroups for phylogenetic analysis.

### Caveats to the study

Two limitations of this study are worthy of note. First, only the relationship between three species is considered; however, as more and more recent divergences become resolved, many phylogenetic problems of central interest increasingly reduce to resolution of the relationship between three or a small number of deeply divergent groups. Second, I only consider the case with one or two outgroups; while this is an unrealistically small number of outgroups for single-gene studies, for genome-wide studies such considerations become more important, particularly given the reduction in numbers of detectable orthologs as more taxa are added.

### Concluding remarks

This discussion of outgroup choice in the presence of differentially retained ancestral gene duplicates yields two arguments supporting the use of closely-related outgroups: (i) for single outgroups, a maximally closely-related post-duplication outgroup either as good as or better than more distant outgroups at reducing non-species tree signal; and (ii) for two outgroups are used, non-species tree signal is minimized when one outgroup is maximally closely-related to the in groups, while the second outgroup position matters less. On the other hand, while the position of the second outgroup position is less important, in the case of different probabilities of retention among duplicates, power to detect gene trees that conflict with known species relationships is maximized when the second outgroup is coincident with or before the duplication event. These findings extend the utility of closely-related outgroups to reducing conflicting signals arising from gene duplication under certain conditions.

## Author Contributions

Conceived and designed the experiments: SWR. Performed the experiments: SWR. Analyzed the data: SWR. Wrote the paper: SWR.

## References

1. Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53: 571–581.
2. Mossel E, Vigoda E (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 30: 2207–2209.
3. Penny D, White WT, Hendy MD, Phillips MJ (2008) A bias in ML estimates of branch lengths in the presence of multiple signals. Mol Biol Evol 25: 239–242.
4. Leigh JW, Susko E, Baumgartner M, Roger AJ (2008) Testing congruence in phylogenomic analysis. Syst Biol 57: 104–115.
5. Huelsenbeck JP, Bull JJ (1996) A likelihood ratio test to detect conflicting phylogenetic signal. Syst Biol 45: 92–98.
6. Bryant D, Moulton V (2002) NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. LCNS 242: 375–391.
7. Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14: 68–73.
8. Fares MA, Byrne KP, Wolfe KH (2006) Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. Mol Biol Evol 23: 245–253.
9. Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate evolution. Genome Biol 8: R141.
10. Zheng C, Zhu Q, Sankoff D (2008) Descendants of whole genome duplication within gene order phylogeny. J. Comput Biol 15: 947–964.
11. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19: 99–113.
12. Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 28: 132–163.
13. Maddison WP (1997) Gene trees in species trees. Syst Biol 46: 523–536.
14. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440: 341–345.
15. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci USA 104: 8397–9402.
16. Sémon M, Wolfe KH (2006) Reciprocal gene loss between *Tetraodon* and zebrafish after whole gene duplication in their ancestor. Trends Genet 23: 108–112.
17. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.
18. McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nat Genet 31: 200–204.
19. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
20. Rogozin IB, Thomson K, Csuros M, Carmel L, Koonin EV (2008) Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's low of homologous series. Biol Direct 3: 7.
21. Hendy M, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool 38: 297–309.
22. Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51: 588–598.
23. Philippe H, Lartillot N, Brinkmann H (2005) Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Mol Biol Evol 22: 1246–1253.